Machine Intelligence Research
www.mi-research.net

20(4), August 2023, 539-553
DOI: 10.1007/s11633-022-1398-0

# Federated Learning on Multimodal Data: A Comprehensive Survey

Yi-Ming Lin  Yuan Gao  Mao-Guo Gong  Si-Jia Zhang
Yuan-Qiao Zhang  Zhi-Yuan Li

School of Electronic Engineering, Xidian University, Xi'an 710071, China

**Abstract:** With the growing awareness of data privacy, federated learning (FL) has gained increasing attention in recent years as a major paradigm for training models with privacy protection in mind, which allows building models in a collaborative but private way without exchanging data. However, most FL clients are currently unimodal. With the rise of edge computing, various types of sensors and wearable devices generate a large amount of data from different modalities, which has inspired research efforts in multimodal federated learning (MMFL). In this survey, we explore the area of MMFL to address the fundamental challenges of FL on multimodal data. First, we analyse the key motivations for MMFL. Second, the currently proposed MMFL methods are technically classified according to the modality distributions and modality annotations in MMFL. Then, we discuss the datasets and application scenarios of MMFL. Finally, we highlight the limitations and challenges of MMFL and provide insights and methods for future research.

**Keywords:** Federated learning, multimodal learning, heterogeneous data, edge computing, collaborative learning.

**Citation:** Y. M. Lin, Y. Gao, M. G. Gong, S. J. Zhang, Y. Q. Zhang, Z. Y. Li. Federated learning on multimodal data: A comprehensive survey. *Machine Intelligence Research*, vol.20, no.4, pp.539–553, 2023. http://doi.org/10.1007/s11633-022-1398-0

## 1 Introduction

With the development of big data, the field of artificial intelligence is increasingly prosperous. The proliferation of edge devices in modern society, as mobile devices and Internet of Things devices, has led to a rapid increase in private data from distributed sources. While abundant data present huge opportunities for solving various tasks, most of the data are not centralized but distributed on the servers of various enterprises, which are highly sensitive in nature. The low-quality, incomplete and insufficient data among multiple parties results in data silos. This is additionally important in the healthcare sector, where medical data are highly sensitive. The data are often kept in different healthcare facilities and are not publicly available[1–4]. As increasing attention is drawn to privacy protection, many laws and regulations such as the General Data Protection Regulation[5], China's Cyber Security Law of the People's Republic of China[6] and the General Principles of the Civil Law of the People's Republic of China[7] have been proposed to protect users' privacy and data security.

Federated learning (FL)[8–11] has received extensive attention as a major paradigm for training models in a privacy-preserving manner. Compared with centralized learning, FL proposes a distributed learning framework that does not transmit private data of each client but only model parameters or intermediate results, which protects the privacy among clients while saving communication costs. The core idea is to perform distributed model training among multiple data sources with local data. The global model is constructed by exchanging model parameters or intermediate results, thus achieving a balance between privacy and accuracy. By eliminating the need to aggregate all data on a single device, FL overcomes the challenges of the privacy concerns mentioned above and allows models to learn from decentralized data.

While FL enables the training of global models without sharing local data, most of the existing FL approaches are trained using unimodal data. With the continuous development of edge computing, various types of sensors and devices generate data from different modalities (e.g., sensory, visual, audio, etc.)[12]. For example, in the medical field, physical examinations usually include X-ray, CT, and MRI; in a smart home, human activities may be recorded by body sensors or by RGB cameras in the room. For clients with different device settings, some of them may have multimodal local data (i.e., multimodal clients) while others may have unimodal local data (i.e., unimodal clients). Therefore, the study of multimodal federated learning (MMFL) is necessary.

Review
Manuscript received on August 19, 2022; accepted on November 25, 2022; published online on June 1, 2023
Recommended by Associate Editor Ji-Rong Wen
Colored figures are available in the online version at https://link.springer.com/journal/11633

Multimodal learning provides higher accuracy and robustness than unimodal learning because multimodal learning combines information from multiple sources at the signal level or semantic level. It better accomplishes the tasks by means of representation[13–15], alignment[16–18], and fusion[19, 20]. Representation finds the contextual properties that reflect the given task by analysing the additional knowledge provided by different modalities. Alignment identifies the mapping between modalities, while fusion combines information from multiple modalities to achieve a prediction task. These methods facilitate FL to use the data uploaded by each client more effectively to accomplish the task, and utilizing these different modalities of client data is the key to achieving better results in FL. Real-world applications of multimodality such as mental health monitoring using wearable sensors[21], medical imaging with RGB and depth maps[22], and language translation using text and images[23] have all shown better accuracy and robustness than unimodal models.

MMFL investigates the problem of learning a global model when each client's local data are of different modalities, which makes it a dark horse in the field of FL for processing multimodal data. MMFL can realize the effective use of complementary information between different modalities and obtain a global model that is superior to unimodal data. This paper is dedicated to a thorough study of current MMFL, including recent advances, challenges and applications. We believe this is the first relatively detailed survey of MMFL. Our main contributions are summarized as follows:

1) We provide a brief overview of FL and analyse in detail the shortcomings of existing FL algorithms when faced with multimodal data. We then introduce the key motivations and application scenarios for using multimodal data for FL. Through these, the proposed taxonomy for MMFL is introduced, and the existing work is introduced and summarized in detail.

2) We propose a modality distribution-based and modality annotation-based classification method for MMFL to showcase existing work on MMFL, highlighting their challenges, their main ideas and assumptions.

3) We introduce MMFL datasets and usage scenarios and envision future MMFL research efforts, and discuss new data-combination methods and more trustworthy methods to build MMFL systems.

The rest of the paper is structured as follows: Section 2 presents related work, including the basics of FL, multimodal learning, and the vision of using multimodal data in FL. The proposed taxonomy is presented in Section 3. In Section 4, some commonly used multimodal federated datasets are introduced. In Section 5, the application of MMFL in reality is given. The current difficulties of MMFL and future perspectives are presented in Section 6. Finally, the paper concludes with the conclusion in Section 7.

## 2 Related works

With the growing interest in privacy-preserving and multimodal data, FL and multimodal learning have been proposed and increasingly applied in real-life applications. In this section, we will introduce the basic concepts of FL and multimodal learning and will give the visions of multimodal data in FL accordingly. This section is the basis for our systematic and comprehensive survey of MMFL.

### 2.1 Multimodal learning

In the context of human-computer interaction, a modality is the classification of a single independent channel of sensory input/output between a computer and a human[24]. Devices such as cameras and microphones directly act on human senses of the computing world, while devices such as sensors are indirect. Multimodality combines multiple modalities to help us better understand the world around us and provide a better experience[25]. With the popularization of smart devices and the continuous progress of deep learning, research questions naturally point to the problem of larger and more complex multimodal data. Multimodal learning can learn the characteristics of data representation at different levels of abstraction and extract useful features from multiple modalities, which makes it more attractive when dealing with different types of data. For example, Bayoudh et al.[26] focused on computer vision-related fields such as vision and language, summarizing six perspectives in the current literature on deep multimodal learning. Then this paper surveys current multimodal applications and proposes a set of benchmark datasets to address problems in various vision domains. Gao et al.[27] mainly explored the fusion field in the multimodal challenge, summarized the current pioneering multimodal fusion model, and introduced some challenges and future topics faced by the model. Muhammad et al.[28] aimed to fill the gap in comprehensive research on multimodal learning in the field of smart healthcare, outlining existing multimodal signal fusion and device fusion schemes, different fusion strategies, and the importance of security and privacy. These reviews all demonstrate in detail the superior performance of multimodal learning. As a method of collaborative learning, FL has also aroused the deep interest of researchers in the field of multimodality.

### 2.2 Federated learning

Since it was proposed as a privacy-preserving machine learning paradigm, FL has made multimodal learning more secure by providing a distributed and privacy-enhanced efficient scheme. With recent advances in mobile hardware and growing concerns about privacy leaks, FL is particularly attractive for building distributed multimodal systems. Homomorphic encryption and other

methods are introduced in information aggregation to avoid violating individual privacy. Thus, user data will only be kept locally, which is beneficial to each participant in the multimodal system in terms of saving communication costs and protecting privacy. Here, we introduce the key concepts of FL in multimodal systems and present some FL categories that are frequently used in multimodal learning.

#### 2.2.1 Key concepts

In a multimodal system, client data for FL can come from a single modality or multiple modalities, such as cameras and sensors in the same scene. The client trains the model locally to extract shared or correlated representations between different modalities, and the original data are kept locally on the client. The server utilizes algorithms such as FedAvg[10] to combine encrypted information from multiple sources to build a better performing global model[9, 29]. The objective function $F(w)$ of the central server is usually expressed as

$$F(w) = \sum_{k=1}^{m} \frac{n_k}{n} F_k(w) \tag{1}$$

where $m$ is the total number of clients involved in training, $n = n_1, \cdots, n_k$ is unimodal or multimodal data and $F_k(w)$ is the local objective function of the $k$-th device,

$$F_k(w) = \frac{1}{n_k} \sum_{i \epsilon d_k} f(x_i, y_i; w) \tag{2}$$

where $d_k$ is the local dataset of the $k$-th client and $f(x_i, y_i; w)$ is the loss function generated by the model with parameter $w$ for the instances $(x_i, y_i)$ in dataset $d_k$.

The sum of the loss functions generated by all instances in $d_k$ divided by the total amount of data in client $k$ is the average loss function of the local clients, and the loss function is inversely proportional to the model accuracy. Therefore, objective function optimization is usually used to make the loss function reach the minimum value.

#### 2.2.2 FL classifications

Depending on how the data are distributed in the feature and sample space, FL frequently used in multimodal systems can be divided into horizontal FL, vertical FL, and federated transfer learning[30].

Horizontal FL is suitable for scenarios in which data holders have the same feature space and little or no overlap in the sample ID space. In a multimodal system, one application of horizontal FL is intelligent monitoring, such as using monitoring or sensors (similar feature space) for people flow detection in different scenarios (sample ID space does not overlap), and the model parameters are averaged through the server to obtain a more efficient global model.

Vertical FL is suitable for scenarios where there is considerable overlap in user space among participants, and little or no overlap in feature space. An example applied in a multimodal system can be e-health, such as the same patient in different hospitals. In this case, the diagnosis results of different parts (different feature spaces) of the same patient (the same sample ID space) are passed through the server by performing parameter alignment, and the patient's electronic medical record can be obtained.

Federated transfer learning is a supplement to horizontal and vertical settings, and is suitable for scenarios where the user space and feature space of each participant have less overlap. The most commonly used scenario for federated transfer learning in the multimodal systems is federated medical care. Many hospitals in different countries collect various medical images and electronic medical records (different feature spaces) from different patients (sample ID spaces do not overlap), so as to effectively improve the model's performance.

### 2.3 Visions of multimodal data in FL

With the emergence of the Internet of Things and the popularization of cloud computing, the rate of generation of multimodal data from heterogeneous sources continues to grow. As a common method for edge computing, traditional FL faces the challenge of diverse client data distributions. In this context, multimodal learning has gained extensive attention from researchers due to its ability to learn deep features from different data and has become an effective method to address data diversity. Through methods such as multimodal representation and fusion, multimodal learning can effectively solve the abovementioned challenges. Therefore, many investigations on FL have focused on areas related to multiple modalities[31–34]. For example, the Internet of Things and big data.

Nguyen et al.[35] provided a comprehensive survey of FL applications, which explores FL and Internet of Things development, integration, and applications. Key challenges such as FL performance, threats, and Internet of Things heterogeneity are analysed in detail, and possible research directions are given. Gadekallun et al.[36] conducted a comprehensive survey of big data services and applications and reviewed the applications and key projects of FL in big data services. The use of FL accelerates the learning process and improves the accuracy of learning while saving communication costs. In addition, FL is also combined with multimodal learning in the medical field. The work in [37] proposed a new dynamic fusion-based FL method for medical diagnostic image analysis to detect COVID-19 infection, which dynamically fuses the client's X-ray and CT data to improve the accuracy of diagnosis on the basis of protecting privacy.

According to the above surveys, MMFL can effectively utilize the information of different modalities to realize the alignment and fusion of modalities and solve

some problems of multimodal learning. In the following sections, we conduct an extensive survey of existing MMFL work and propose a taxonomy. This paper discusses in detail the role and benefits of MMFL and highlights gaps gained from the survey and research for future work.

## 3 Proposed taxonomy

A current challenge of FL research is that there is not a detailed taxonomy of using multimodal data. Many of the papers discussed above, although they focus on MMFL work, only describe its application in a specific field and do not integrate and classify MMFL work. Ideally, the multimodal data held by clients of MMFL are homogeneous and well annotated. However, in real life, multimodal data have distribution diversity, which leads to data heterogeneity among different clients. The data owner does not have the time or ability to label the data.

In this paper, we investigate the question of whether client-side multimodal data are heterogeneous and whether annotated data are sufficient, introducing a new MMFL taxonomy for the first time. Our proposed classification method, shown in Fig. 1, is a key consideration for MMFL based on modal conditions when creating an effective global model. These conditions are the distribution of modalities among clients and the availability of modal annotations. The taxonomy is based on two aspects: modality distributions and modality annotations. For each aspect, we further provide multiple subcategories that reflect the impact of client-side multimodal data and labels on MMFL.

### 3.1 Modality distributions

The client data of FL are generally unimodal and homogeneous (Fig. 2(a)), which allows the aggregation server to directly use algorithms such as federated averaging. However, with the popularization of smart devices and the increasing demand for deep learning, client-side data will have modal heterogeneity, and traditional FL can no longer meet the needs of realistic requirements. For example, in a smart home, surveillance video or motion sensor data belong to different modalities that describe the same thing, and either of these can be used to complete the gesture recognition task, but FL cannot directly use the uploaded data of the two to build a global model. How to align data from different modalities, how to take advantage of complementary information between modalities, and how to preprocess client-side modalities to generate a global model are all issues that this subsection focuses on. Starting from the modalities of the client, we design the basic model for MMFL in Fig. 2(b) and find through analysis that MMFLs are generally divided into two types: 1) homogeneous multimodality and 2) heterogeneous hybrid modality. A schematic diagram of these two types is shown in Fig. 3. It is worth noting that these works tend to be supervised learning, and related semi-supervised and unsupervised work will be discussed on modality annotations classification.

#### 3.1.1 Homogeneous multimodality

We start our discussion with homogeneous multimodality, which, in simple terms, replaces the unimodal client in FL with a client that uses multimodal data. Homogeneous multimodality is an initial successful attempt to combine FL and multimodal learning, and is mainly used for tasks where centralized learning cannot be achieved due to privacy or communication cost constraints[38]. These tasks are often limited by the number of samples and privacy, and the performance of locally trained models is poor and insufficient to meet the task requirements. With the help of the FL structure, the work in homogeneous multimodality overcomes the limitation of the number of samples and achieves the performance improvement of each client.

Agbley et al.[39] fused two modalities: skin lesion images and their corresponding clinical data. By comparing the results of centralized learning and FL, it is demonstrated that MMFL can effectively learn highly predictive models while ensuring that training data are not shared among participating clients. With the successful application of MMFL in the medical field, relevant researchers have set their sights on the field of autonomous driving. Cassará et al. [40] proposed a new federated fea-

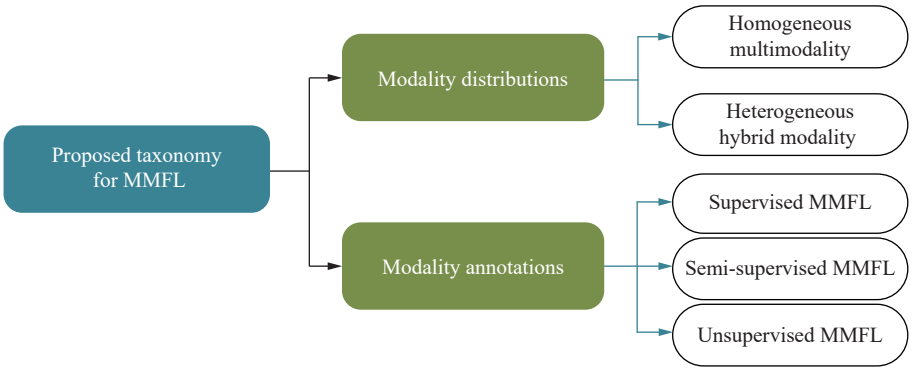
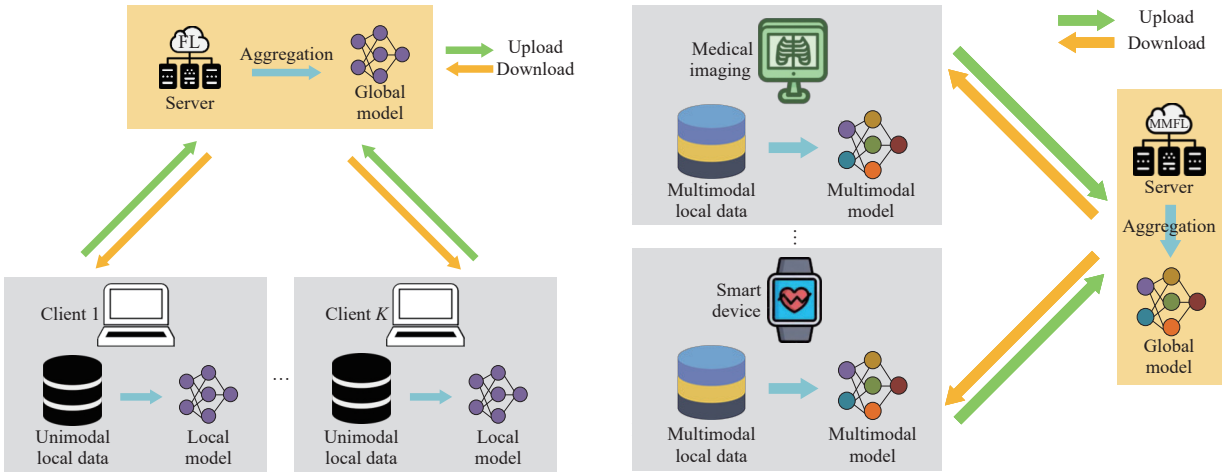
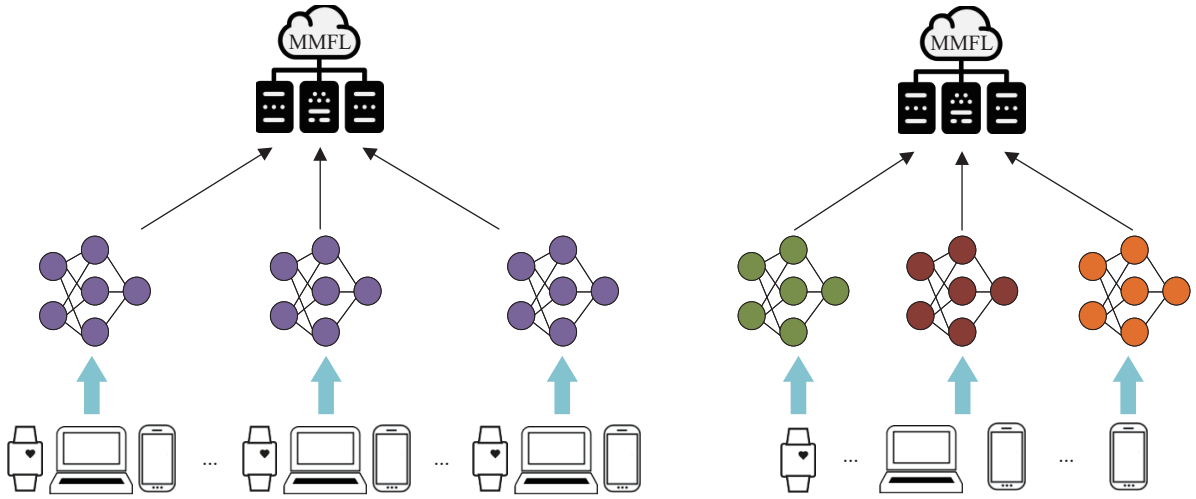



Fig. 1 Proposed taxonomy for MMFL




(a) Federated learning (b) Multimodal federated learning

Fig. 2 The basic model of federated learning and multimodal federated learning: (a) Federated learning, which trains a global model by integrating training parameters from each unimodal client under the control of a central server; (b) Multimodal federated learning that enhances the accuracy and robustness of FL.




(a) Homogeneous multimodality (b) Heterogeneous hybrid modality

Fig. 3 A schematic diagram of the modality distributions classification model: (a) Homogeneous multimodality, where the local data are multimodal and homogeneous between clients; (b) Heterogeneous hybrid modality, where local data are a hybrid modality that is heterogeneous between clients.

ture selection algorithm that iteratively proposes the smallest set of features selected from the local dataset. Feature selection is carried out using the mutual information metric, and the optimization problem is solved by the method of cross entropy. The algorithm uses a Bayesian-based aggregation method to send the client's information to the edge server. In the same period, Salehi et al.[41] proposed a multimodal FL framework that used multimodal data from sensors and millimeter radar to predict optimal sectors. They also investigated the impact of missing modalities during inference. The results show that using the proposed MMFL framework reduces sector selection time by 52% while maintaining sufficient throughput.

It can be seen from the above investigation that homogeneous multimodal work can effectively solve the problems of insufficient samples and high communication costs, which is a successful attempt at FL in the multimodal field. However, this classification work also has great shortcomings; that is, it requires homogeneous data between clients and cannot effectively deal with the problems of heterogeneous modalities. Then, we discuss the case where data on the client side are heterogeneous.

### 3.1.2 Heterogeneous hybrid modality

One of the main challenges faced by MMFL is that multimodal data in real-world scenarios often suffer from heterogeneous modalities. Clients in this category have data in one or more modalities, and there is data heterogeneity[42–45] among them. The main reason for this is the heterogeneity between devices, such as hospitals in first-

tier cities and hospitals in remote areas in smart healthcare: hospitals in first-tier cities may use CT, X-ray and other methods to detect the same part and use EHR as an aid, while remote areas may only have one detection method. To solve the heterogeneous problem, MMFL utilizes alignment, fusion and collaborative learning[46–48] methods to effectively use heterogeneous data to build a global model and improve the performance of each participant model.

We study a recent paper on modality alignment[49], and the method in the paper does not achieve modality alignment but rather early fusion. Fused training parameters cannot be applied to unimodal clients, to capture the complementarity of information between different modalities in a distributed manner, Xiong et al.[50] proposed a simple and effective multimodal FL framework. The attention mechanism, as a common method in multimodal alignment methods, is used in the framework to capture the complementary information between each client′s sensor data and images. Wei[51] applied FL to a multimodal fusion model and proposed a new algorithm for FL-based disease diagnosis models. The client side independently trains the classification model for each modal data and uploads the parameters, and then the server side applies an aggregation algorithm to analyse their gradient descent direction. Qayyum et al.[52] proposed a collaborative learning framework based on clustering FL to intelligently process visual data at the edge by training a multimodal learning model capable of diagnosing COVID-19 in X-ray and ultrasound images. With collaborative learning, clustering FL can handle heterogeneous data from different sources (i.e., X-ray and ultrasound images).

The work in heterogeneous hybrid modality solves the problem of heterogeneous modalities caused by the diversity of data distribution with the help of multimodality methods. In addition, Bernecker et al.[53] normalized the data with the help of slice modality information and hard-coded mode assignment, and achieved high-performance completion of liver segmentation tasks for multimodal data. However, compared with homogeneous multimodality, the above method consumes more computing resources and has the disadvantages of relying on complementary information between modalities and high time costs.

#### 3.1.3 Discussion

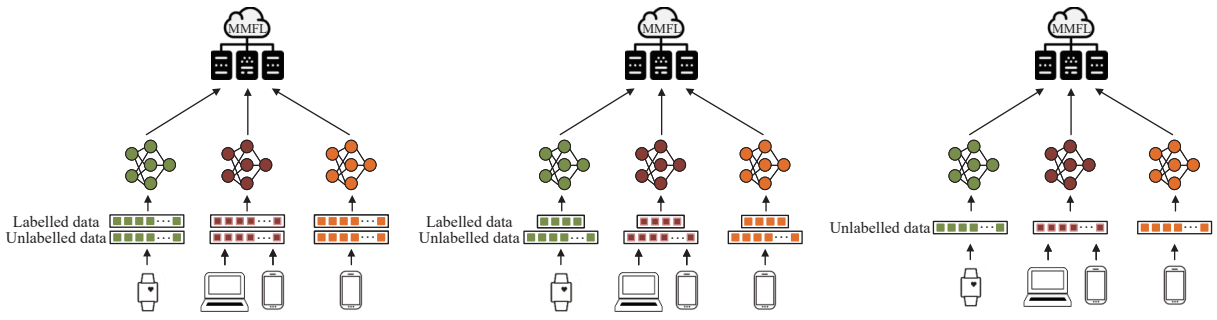
In this subsection, we propose two classifications based on modality distributions: homogeneous multimodality and heterogeneous hybrid modality. Homogeneous multimodality follows the structure of FL, replaces unimodal clients with multimodal clients, and makes use of the advantages of FL′s network structure to compensate for the lack of local sample size. It is most suitable for the case where clients are homogeneous to each other. It is widely used in autonomous driving and smart homes. Heterogeneous hybrid modality takes into account the data heterogeneity caused by device heterogeneity in practical applications and can effectively aggregate various clients. However, both categories default to all local client data being annotated, which is difficult to achieve in an information explosion environment. Next, we present the work of MMFL under different annotation conditions.

### 3.2 Modality annotations

The continuous development of big data makes it very easy for people to obtain data, which provides data support for the development of multimodal learning. However, the owners of these data may not have the time or ability to label the data, such as judging whether the CT or X-ray images of an organ are normal. Annotation via user interaction is also not possible, as this would lead to a risk of leakage of user privacy. This subsection looks at modal annotation in MMFL, summarizing and analysing supervised work, semi-supervised work[54–61] and unsupervised work[62–66]. The schematic diagram of modality annotations is shown in Fig. 4.

#### 3.2.1 Supervised MMFL

In supervised models, all samples of all modalities have labels available. The initial implementation of MMFL started with supervised learning on labelled data. With the help of FL, a multimodal transfer learning framework is studied in [67] to build a powerful global




(a) Supervised MMFL (b) Semi-supervised MMFL (c) Unsupervised MMFL

Fig. 4 A schematic diagram of the modality annotations classification model: (a) Supervised MMFL, trained with labelled data; (b) Semi-supervised MMFL, trained with both labelled and unlabelled data; (c) Unsupervised MMFL, trained with unlabelled data.

model by aggregating labelled data from different hospital organizatiosns with the help of multiple smart wearable devices.

The disease diagnostic multimodal model[52] uses fully labelled CT and X-ray images for alignment to exploit the complementary information of multiple modalities. Chen and Zhang[68] proposed the FedMSplit framework, which employs a dynamic graph structure to capture adaptive correlations between multimodal client models. The framework conducts experiments using a labelled multimodal ensemble dataset, demonstrating the effectiveness of the method for MMFL problems with inconsistent modalities among clients.

Labelled data allow researchers to explore the relationship between features and labels to quickly build corresponding models. However, sufficient labelled data do not exist in real life, and the semi-supervised and unsupervised MMFL described below is trained using unlabelled data.

#### 3.2.2 Semi-supervised MMFL

For multimodal models, two situations may occur with the client data of semi-supervised MMFL: All clients have a small amount of labelled data and a large amount of unlabelled data; some clients use labelled data, and others use unlabelled data. The number of modalities is different, and the number of combinations produced is also different, such as multimodal learning using labeled images as well as large amounts of unlabelled text data[69, 70]. Although partially labelled data exist for semi-supervised learning, labelling data is a time-consuming manual process and prone to errors. Therefore, it is necessary to use semi-supervised MMFL methods to achieve effective utilization of multimodal data. One of the most commonly used methods in semi-supervised is the autoencoder[71–73] method, which enables representation learning on the input data.

Zhao et al.[74] proposed a semi-supervised FL framework for human activity recognition and compared the performance of different autoencoders. Their framework has better performance than data augmentation schemes. On this basis, a multimodal and semi-supervised FL framework is proposed in [75] and assumes that unlabelled data can come from a unimodal modality or from multiple modalities. The framework trains an autoencoder to extract shared or correlated representations from different local data patterns on the client side, and performs downstream classification on auxiliary labelled data via a global autoencoder. In addition, Yu et al.[76] proposed a federated framework based on semi-supervised online learning, which introduces an aligned hierarchical attention architecture for different levels of features, which can achieve better results through unsupervised gradient aggregation and fine-tuning.

#### 3.2.3 Unsupervised MMFL

In unsupervised models, the training data are not labelled. This is because labelling these data is expensive and challenging. Multimodality makes this step extra difficult, as each modality requires individual annotations and some annotations require expert annotation[77]. MMFL can use unannotated multimodal data to accomplish the task in several ways. Data augmentation is a method to achieve unsupervised MMFL. Xie et al.[78] designed a data augmentation procedure for unsupervised training of FL using severely distorted and unpaired data to complete multimodal brain synthesis.

Because of privacy, bandwidth and high annotation cost, Saeed et al.[79] proposed an unsupervised method for scalogram-signal correspondence learning. The method learns useful representations from unlabelled sensor inputs by designing auxiliary tasks. To overcome the limitation of huge memory and computational cost required to directly train unlabelled data in smart healthcare, Arikumar et al.[80] proposed an FL-based person movement identification algorithm. Using a deep reinforcement learning framework to automatically label unlabelled data achieves high accuracy while reducing computational cost and memory usage. A federated transfer learning framework was proposed in [81], which uses contrastive learning as a transfer learning strategy to learn transferable representations from multimodal participants to help participants work with unimodal data.

Therefore, utilizing unsupervised techniques to learn useful representations is important for unlabelled multimodal data. The above methods show good results on the MMFL task.

#### 3.2.4 Discussion

Supervised MMFL using annotated data can effectively address the problem of heterogeneous modalities and has higher robustness. Nevertheless, adding annotations manually is time-consuming, and some annotations require expert knowledge. The annotation of multimodal data is more complex, and it is necessary to ensure the alignment of annotations and modalities. Therefore, work on semi-supervised and unsupervised MMFL is necessary and important.

## 4 Multimodal federated datasets

The availability of multisource data from various sensors and mobile devices drives the trend of multimodal data usage, and MMFL works by slicing multimodal datasets to represent clients that are homogeneous or heterogeneous to each other. To date, a series of multimodal datasets for various aspects has been available, which has greatly facilitated the development of MMFL. Faced with such a rich multimodal dataset, which data should be selected for MMFL research is a matter of careful consideration. To this end, in this section, we introduce several commonly used multimodal datasets in MMFL work, including domains such as autonomous driving, object recognition, and medical diagnosis.

**Kineics-400**[82]**.** It consists of a massive dataset of

YouTube video URLs, which includes a range of different human actions. The dataset includes more than 300 000 video sequences of 400 human actions.

**RealWorld2**[83]**.** It is a sensor-based HAR dataset collected from 15 probands. It includes 8 activities, such as running, standing and lying down. Everyone wears sensing devices on seven parts of the body, including the chest, forearm, head, calf, thigh, upper arm and waist.

**IEMOCAP**[84]**.** It consists of 4 453 video clips of recorded conversations. Each segment is annotated with 9 emotions (happy, angry, excited, fearful, etc.) for human emotion analysis.

**ModelNet40**[85]**.** It contains 12 311 3D shapes that are used for multiview 3D object recognition tasks, covering 40 common categories, including airplanes, bathtubs, beds, bookshelves, and more. Every 3D CAD object has an $M = 2$ mode as two views of its shape.

**Vehicle sensor**[86]**.** It contains 23 instances. Each instance is an individual client described by 50 acoustic features and 50 seismic features. This dataset is often used for vehicle identification.

**mHealth dataset**[87]**.** It contains 13 activities of daily living and exercise, including standing still, sitting and relaxing, and zero activity. These activities are measured by multimodal body sensors, including accelerometers, ECG sensors, gyroscopes, and magnetometers.

**UR fall detection dataset**[88]**.** It contains 70 video clips of human activities recorded by an RGB camera and a depth camera, including not lying, lying on the ground, and temporary poses. Each video frame was labelled and paired with sensory data from an accelerometer measured in grams.

# 5 Applications

The excellent performance exhibited by MMFL has led to its integration with real-world applications. Researchers have used MMFL to improve the performance of designed tasks to better accomplish learning goals and meet real-life needs. In this section, we discuss MMFL application scenarios as well as recent state-of-the-art literature to introduce the advantages of MMFL in solving real-life problems in more detail.

## 5.1 MMFL in internet of things

Through various devices and technologies, such as various information sensors, radio frequency identification technology, and infrared sensors, the Internet of Things collect various needed information such as sound, location, and movement, which facilitates human life[89–91]. However, most of the data generated by sensors are inherently sensitive and involve the private data of different citizens and businesses. Modern applications usually deploy different types of sensors or devices, mostly generating data from different modalities (e.g., visual, audio, and other senses). For example, in a human activity recognition task, human behavior can be recorded by an RGB camera or by a sensor for gesture detection[92]. Traditional FL or multimodal learning approaches cannot handle the problems that may arise in the Internet of Things. MMFL, which combines the practicalities of FL and multimodal learning, is a potential solution to the above problems. Recent research on MMFLs used to address Internet of Things problems will be discussed below.

Zhu et al.[93] proposed a new framework based on hybrid policies, while introducing the edge FL model into reinforcement learning and developing a corresponding online algorithm to improve the system performance. The algorithm uses sensor data, urban security surveillance, and data from smart device users for certain types of computing tasks to achieve tasks such as trajectory planning, data scheduling, and bandwidth allocation. The results show that the algorithm not only outperforms the baseline in terms of the average system age but also improves the stability of the training process. Considering practical network resource constraints, Wang et al.[94] proposed a more robust system-level multitask federated connected autonomous vehicle perception. Perceptual errors are minimized by multilayer graph resource allocation and vehicle pose comparison methods. By analysing point clouds, images, and radar data collected by onboard sensors, it also automatically adjusts the number of sensors in the cases of varying complexity, enabling automatic regulation of resources. The algorithms are experimentally proven to greatly improve the perceptual accuracy for all tasks. Chen and Li[95] found the reasons why some MMFL performance lags behind FL by analysing the complexity of non-iid data, namely, the data distribution and modal distribution across clients and its heterogeneity. To address this issue, they proposed a new training algorithm, called hierarchical gradient blending, which adaptively achieves optimal mixing of modal subnetworks and optimal aggregation of local updates. A rigorous theoretical analysis is used to demonstrate the effectiveness in various non-IID multimodal data scenarios.

## 5.2 MMFL in HealthCare

With the rapid development of computer software and hardware technologies, more and more healthcare data are available from organizations such as clinical facilities, patients and the pharmaceutical industry[36, 96]. It has enabled the healthcare field to make better use of all types of data to improve the quality of diagnosis and care. However, because these healthcare data are often decentralized and private, unauthorized and casual use of the data can lead to personal and institutional privacy breaches, creating a huge information security crisis. The diagnosis of a disease is often related to multiple data, which involves the use of multimodal learning, such as

COVID-19, which affects the whole world, can be determined not only from lung CT or X-ray but also through electronic health records[97–99]. This poses a challenge for technologies applied to healthcare, and MMFL, as a technology that enables distributed learning using multimodal data while protecting privacy, is a key approach to this challenge.

Parekh et al.[100] explored FL that contains datasets from different domains and is trained to solve different tasks. The paper evaluates cross-domain FL for target detection and segmentation tasks in two different experimental settings: multimodal and multiorgan. The model achieves tumor segmentation by CT and PET of the kidney as well as MRI of the brain and MRI of the chest. The results demonstrate the potential of FL in developing multidomain, multitask deep learning models without sharing data from different domains. To diagnose COVID-19 effectively without compromising privacy, Chen et al.[101] discussed a multimodal study of COVID-19 based on collaborative joint learning and proposed a blockchain FL model to diagnose COVID-19 by analysing CT and X-ray images. In this model, different hospitals train the parameters using local data and then upload them to the blockchain for encryption, and the whole process protects the privacy of patients. The model has proven to be highly feasible in practice. Ji et al.[102] proposed a vertical FL model for detecting human states using multimodal data. The model considers the speed problem due to device heterogeneity and different modal complexities and designs a fast and secure module that effectively reduces the amount of transmitted data. Experiments using brain CT images and a time-series heart rate dataset for human state detection show the feasibility of multimodal vertical FL. Compared with single-peak learning, the accuracy was improved by 5.6 percentage points, and the time required for training was reduced by 48.1%.

# 6 Research challenges and future directions

With the continuous progress and development of big data, practical MMFL applications have started to demand better models, and this area is attracting increasing attention. Based on our review of MMFL classification and MMFL applications in different domains above, we summarize the difficulties and challenges of existing MMFL and propose foreseen research directions to address these research difficulties.

## 6.1 Multimodal alignment

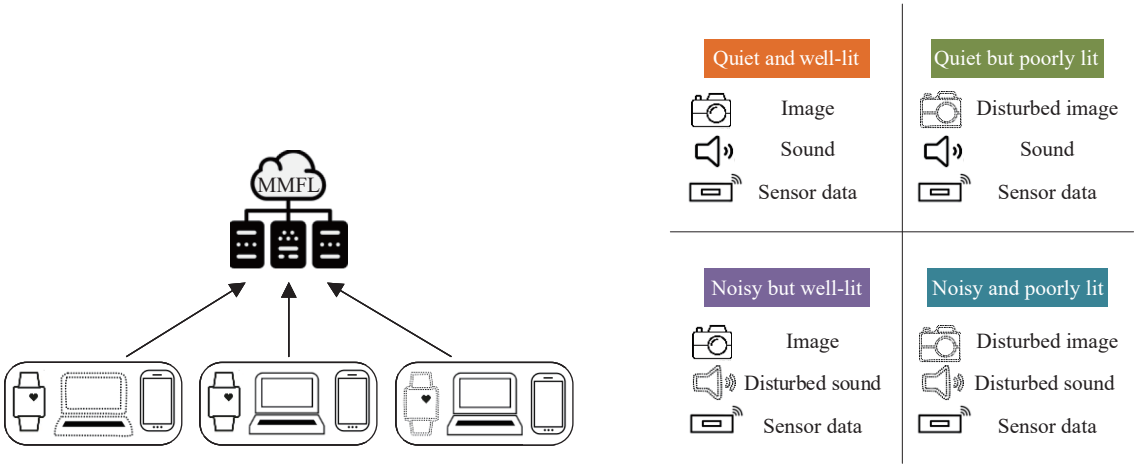
Multimodal alignment is defined as finding relationships and correspondences between components of multiple modal instances. For example, given an image and a paragraph describing the image, we want to find the correspondence between the textual description area and the object of the given image[77, 103, 104]. For MMFL, aligning the signals from different modalities is an important step because it affects the performance of the later stages. Most of the current MMFL models require client-side alignment between single-peak data, while few datasets have explicitly labelled alignments, which are typically created manually. It requires preprocessing of multimodal data or finding common features in higher dimensions of different modal data, which is a limitation for the real-time performance of the algorithm and for reducing the time required to train and use the model. There may be possible alignment, not all elements in one modality have correspondence in another modality. To address the above issues, future research should focus on unsupervised learning that does not require labelled data and self-supervised learning that can be trained by contextual relationships of unlabelled data, or semi-supervised learning by extracting hidden representations of unlabelled data and correlated representations between different modalities.

## 6.2 Modal deficiency

Modal deficiency can be caused by 1) the heterogeneity of devices used by different clients, resulting in one or more missing modes, and 2) the devices being the same but in different scenarios, resulting in a certain modality being interfered with and not being used[105–108]. A schematic diagram of modality deficiency can be seen in Fig. 5. Although the study of overcoming modal deficiency to complete the task has been introduced in Section 3 on modality distribution, most of these studies obtained more generalized results by some unsupervised methods, which is not enough for some tasks requiring higher accuracy. The modal deficiency client has higher computational costs than the modal complete client, which is the opposite of the problem MMFL tries to solve. To solve the modal deficiency, future research can focus on using data augmentation to complement missing modalities or modal normalization to achieve better results by complementing modalities or ignoring the influence between modalities.

## 6.3 Privacy protection

To protect the private data of clients from leakage, existing research protects users′ data security through privacy-preserving approaches such as secure multiparty computation and differential privacy[109, 110]. While MMFL utilizes multimodal data to obtain better robustness and accuracy, the alignment information of local clients themselves and between clients increases the risk of data leakage[111]. Through attacks such as poisoning attacks or inference attacks, hostile participants can infer the private data of other clients by uploading fake data and obtaining additional information with the help of




(a) Different devices (b) Different scenarios

Fig. 5 (a) Device heterogeneity causes modal deficiencies; (b) Different scenarios result in one or more modalities being disturbed and not being used.

alignment information or update parameters. To address this challenge, future research should, on the one hand, continue to strengthen the research on various privacy protection methods related to FL and, on the other hand, consider preprocessing the alignment information to prevent adversarial participants from using this information to infer private information.

## 7 Conclusions

In this survey, we systematically review the strengths as well as the weaknesses of FL and multimodal learning and use them to propose the need for using MMFL. We propose a technical classification of MMFL based on the key challenges in MMFL and existing research, and provide a detailed explanation of the reasons for the classification. Finally, we discuss the applications of MMFL and present the issues and directions for further research on MMFL. We hope that the survey can help relevant researchers understand the current status and importance of MMFL research and contribute to the long-term development of MMFL.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62036006), the Fundamental Research Funds for the Central Universities, China, and the Innovation Fund of Xidian University, China.

## Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

## References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang. Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, ACM, Vienna, Austria, pp. 308–318, 2016. DOI: 10.1145/2976749.2978318.

[2] G. A. Kaissis, M. R. Makowski, D. Rückert, R. F. Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020. DOI: 10.1038/s42256-020-0186-1.

[3] Y. Gao, M. G. Gong, Y. Xie, A. K. Qin, K. Pan, Y. S. Ong. Multiparty dual learning. *IEEE Transactions on Cybernetics*, published online. DOI: 10.1109/TCYB.2021.3139076.

[4] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. L. Xu, D. Marcus, R. R. Colen, S. Bakas. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, vol. 10, no. 1, Article number 12598, 2020. DOI: 10.1038/s41598-020-69250-1.

[5] J. P. Albrecht. How the GDPR will change the world. *European Data Protection Law Review*, vol. 2, no. 3, pp. 287–289, 2016. DOI: 10.21552/EDPL/2016/3/4.

[6] M. Parasol. The impact of China's 2016 cyber security law on foreign technology firms, and on China's big data and smart city dreams. *Computer Law & Security Review*, vol. 34, no. 1, pp. 67–98, 2018. DOI: 10.1016/j.clsr.2017.05.022.

[7] W. Gray, H. R. Zheng. General principles of civil law of the people's republic of China. *The American Journal of Comparative Law*, vol. 34, no. 4, pp. 715–743, 1986. DOI: 10.2307/840330.

[8] M. G. Gong, Y. Xie, K. Pan, K. Y. Feng, A. K. Qin. A survey on differentially private machine learning [Review Article]. *IEEE Computational Intelligence Magazine*, vol. 15, no. 2, pp. 49–64, 2020. DOI: 10.1109/MCI.2020.2976185.

[9] Q. Yang, Y. Liu, T. J. Chen, Y. X. Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, Article number 12, 2019. DOI: 10.1145/3298981.

[10] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A.

Y. Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, USA, pp. 1273–1282, 2017.

[11] Y. Zhao, M. Li, L. Z. Lai, N. Suda, D. Civin, V. Chandra. Federated learning with non-IID data, [Online], Available: https://arxiv.org/abs/1806.00582, 2018.

[12] A. Brunete, E. Gambao, M. Hernando, R. Cedazo. Smart assistive architecture for the integration of iot devices, robotic systems, and multimodal interfaces in healthcare environments. *Sensors*, vol. 21, no. 6, Article number 2212, 2021. DOI: 10.3390/s21062212.

[13] Y. Mroueh, E. Marcheret, V. Goel. Deep multimodal learning for audio-visual speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, South Brisbane, Australia, pp. 2130–2134, 2015. DOI: 10.1109/ICASSP.2015.7178347.

[14] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ACM, Bellevue, USA, pp. 689–696, 2011. DOI: 10.5555/3104482.3104569.

[15] Y. W. Pan, T. Mei, T. Yao, H. Q. Li, Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 4594–4602, 2016. DOI: 10.1109/CVPR.2016.497.

[16] D. Bahdanau, K. Cho, Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, USA, 2015.

[17] A. Karpathy, F. F. Li. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 3128–3137, 2015. DOI: 10.1109/CVPR.2015.7298932.

[18] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, pp. 2048–2057, 2015.

[19] X. Y. Jiang, F. Wu, Y. Zhang, S. L. Tang, W. M. Lu, Y. T. Zhuang. The classification of multi-modal data with hidden conditional random field. *Pattern Recognition Letters*, vol. 51, pp. 63–69, 2015. DOI: 10.1016/j.patrec.2014.08.005.

[20] G. A. Ramirez, T. Baltrušaitis, L. P. Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction*, Springer, Memphis, USA, pp. 396–406, 2011. DOI: 10.1007/978-3-642-24571-8_51.

[21] E. Garcia-Ceja, M. Riegler, T. Nordgreen, P. Jakobsen, K. J. Oedegaard, J. Tøresen. Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing*, vol. 51, pp. 1–26, 2018. DOI: 10.1016/j.pmcj.2018.09.003.

[22] Y. D. Xia, D. Yang, Z. D. Yu, F. Z. Liu, J. Z. Cai, L. Q. Yu, Z. T. Zhu, D. G. Xu, A. Yuille, H. Roth. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Medical Image Analysis*, vol. 65, Article number 101766, 2020. DOI: 10.1016/j.media.2020.101766.

[23] J. Rajendran, M. M. Khapra, S. Chandar, B. Ravindran. Bridge correlational neural networks for multilingual multimodal representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, San Diego, USA, pp. 171–181, 2016. DOI: 10.18653/v1/N16-1021.

[24] F. Karray, M. Alemzadeh, J. A. Saleh, M. N. Arab. Human-computer interaction: Overview on state of the art. *International Journal on Smart Sensing and Intelligent Systems*, vol. 1, no. 1, pp. 137–159, 2008. DOI: 10.21307/ijssis-2017-283.

[25] N. Rieke, J. Hancox, W. Q. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. M. Summers, A. Trask, D. G. Xu, M. Baust, M. J. Cardoso. The future of digital health with federated learning. *npj Digital Medicine*, vol. 3, Article number 119, 2020. DOI: 10.1038/s41746-020-00323-1.

[26] K. Bayoudh, R. Knani, F. Hamdaoui, A. Mtibaa. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *The Visual Computer*, vol. 38, no. 8, pp. 2939–2970, 2022. DOI: 10.1007/s00371-021-02166-7.

[27] J. Gao, P. Li, Z. K. Chen, J. N. Zhang. A survey on deep learning for multimodal data fusion. *Neural Computation*, vol. 32, no. 5, pp. 829–864, 2020. DOI: 10.1162/neco_a_01273.

[28] G. Muhammad, F. Alshehri, F. Karray, A. E. Saddik, M. Alsulaiman, T. H. Falk. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Information Fusion*, vol. 76, pp. 355–375, 2021. DOI: 10.1016/j.inffus.2021.06.007.

[29] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'oliveira, H. Eichner, S. El Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. Y. He, L. He, Z. Y. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, H. Qi, D. Ramage, R. Raskar, M. Raykova, D. Song, W. K. Song, S. U. Stich, Z. T. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Y. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, S. Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, vol. 14, no. 1-2, pp. 1–210, 2021. DOI: 10.1561/2200000083.

[30] T. Li, A. K. Sahu, A. Talwalkar, V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020. DOI: 10.1109/MSP.2020.2975749.

[31] A. M. Fu, X. L. Zhang, N. X. Xiong, Y. S. Gao, H. Q. Wang, J. Zhang. VFL: A verifiable federated learning with privacy-preserving for big data in industrial IoT. *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 3316–3326, 2022. DOI: 10.1109/TII.2020.3036166.

[32] B. Zhao, K. Fan, K. Yang, Z. L. Wang, H. Li, Y. T. Yang. Anonymous and privacy-preserving federated learning with industrial big data. *IEEE Transactions on Industrial Informatics*, vol. 17, no. 9, pp. 6314–6323, 2021. DOI: 10.1109/TII.2021.3052183.

[33] Y. L. Lu, X. H. Huang, Y. Y. Dai, S. Maharjan, Y. Zhang. Blockchain and federated learning for privacy-preserved data sharing in industrial IoT. *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4177–4186, 2020. DOI: 10.1109/TII.2019.2942190.

[34] I. Kholod, E. Yanaki, D. Fomichev, E. Shalugin, E. Novikova, E. Filippov, M. Nordlund. Open-source federated learning frameworks for IoT: A comparative review and analysis. *Sensors*, vol. 21, no. 1, Article number 167, 2020. DOI: 10.3390/s21010167.

[35] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, H. V. Poor. Federated learning for internet of things: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021. DOI: 10.1109/COMST.2021.3075439.

[36] S. Huang, W. Shao, M. L. Wang, D. Q. Zhang. fMRI-based decoding of visual information from human brain activity: A brief review. *International Journal of Automation and Computing*, vol. 18, no. 2, pp. 170–184, 2021. DOI: 10.1007/s11633-020-1263-y.

[37] W. S. Zhang, T. Zhou, Q. H. Lu, X. Wang, C. S. Zhu, H. Y. Sun, Z. P. Wang, S. K. Lo, F. Y. Wang. Dynamic-fusion-based federated learning for COVID-19 detection. *IEEE Internet of Things Journal*, vol. 8, no. 21, pp. 15884–15891, 2021. DOI: 10.1109/JIOT.2021.3056185.

[38] A. Nandi, F. Xhafa. A federated learning method for real-time emotion state classification from multi-modal streaming. *Methods*, vol. 204, pp. 340–347, 2022. DOI: 10.1016/j.ymeth.2022.03.005.

[39] B. L. Y. Agbley, J. P. Li, A. U. Haq, E. K. Bankas, S. Ahmad, I. O. Agyemang, D. Kulevome, W. D. Ndiaye, B. Cobbinah, S. Latipova. Multimodal melanoma detection with federated learning. In *Proceedings of the 18th International Computer Conference on Wavelet Active Media Technology and Information Processing*, IEEE, Chengdu, China, pp. 238–244, 2021. DOI: 10.1109/ICCWAMTIP53232.2021.9674116.

[40] P. Cassará, A. Gotta, L. Valerio. Federated feature selection for cyber-physical systems of systems. *IEEE Transactions on Vehicular Technology*, vol. 71, no. 9, pp. 9937–9950, 2022. DOI: 10.1109/TVT.2022.3178612.

[41] B. Salehi, J. Gu, D. Roy, K. Chowdhury. FLASH: Federated learning for automated selection of high-band mmWave sectors. In *Proceedings of IEEE Conference on Computer Communications*, IEEE, London, UK, pp. 1719–1728, 2022. DOI: 10.1109/INFOCOM48880.2022.9796865.

[42] D. L. Li, J. P. Wang. FedMD: Heterogenous federated learning via model distillation, [Online], Available: https://arxiv.org/abs/1910.03581, 2019.

[43] M. S. H. Abad, E. Ozfatura, D. GUndUz, O. Ercetin. Hierarchical federated learning ACROSS heterogeneous cellular networks. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Barcelona, Spain, pp. 8866–8870, 2020. DOI: 10.1109/ICASSP40776.2020.9054634.

[44] A. Khaled, K. Mishchenko, P. Richtárik. Tighter theory for local SGD on identical and heterogeneous data. *Proceedings of Machine Learning Research*, vol. 108, pp. 4519–4529, 2020.

[45] Q. L. Dang, W. Xu, Y. F. Yuan. A dynamic resource allocation strategy with reinforcement learning for multimodal multi-objective optimization. *Machine Intelligence Research*, vol. 19, no. 2, pp. 138–152, 2022. DOI: 10.1007/s11633-022-1314-7.

[46] D. Spikol, E. Ruffaldi, L. Landolfi, M. Cukurova. Estimation of success in collaborative learning based on multimodal learning analytics features. In *Proceedings of the 17th International Conference on Advanced Learning Technologies*, IEEE, Timisoara, Romania, pp. 269–273, 2017. DOI: 10.1109/ICALT.2017.122.

[47] J. K. Olsen, K. Sharma, N. Rummel, V. Aleven. Temporal analysis of multimodal data to predict collaborative learning outcomes. *British Journal of Educational Technology*, vol. 51, no. 5, pp. 1527–1547, 2020. DOI: 10.1111/bjet.12982.

[48] W. X. Hu, B. Cai, A. Y. Zhang, V. D. Calhoun, Y. P. Wang. Deep collaborative learning with application to the study of multimodal brain development. *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 12, pp. 3346–3359, 2019. DOI: 10.1109/TBME.2019.2904301.

[49] P. P. Liang, R. Salakhutdinov, L. P. Morency. Computational modeling of human multimodal language: The MOSEI dataset and interpretable dynamic fusion. In *the 1st Workshop and Grand Challenge on Computational Modeling of Human Multimodal Language*, 2018.

[50] B. C. Xiong, X. S. Yang, F. Qi, C. S. Xu. A unified framework for multi-modal federated learning. *Neurocomputing*, vol. 480, pp. 110–118, 2022. DOI: 10.1016/j.neucom.2022.01.063.

[51] X. Y. Wei. A multi-modal heterogeneous data mining algorithm using federated learning. *The Journal of Engineering*, vol. 2021, no. 8, pp. 458–466, 2021. DOI: 10.1049/tje2.12049.

[52] A. Qayyum, K. Ahmad, M. A. Ahsan, A. Al-Fuqaha, J. Qadir. Collaborative federated learning for healthcare: Multi-modal COVID-19 diagnosis at the edge. *IEEE Open Journal of the Computer Society*, vol. 3, pp. 172–184, 2022. DOI: 10.1109/OJCS.2022.3206407.

[53] T. Bernecker, A. Peters, C. L. Schlett, F. Bamberg, F. Theis, D. Rueckert, J. Weiß, S. Albarqouni. FedNorm: Modality-based normalization in federated learning for multi-modal liver segmentation, [Online], Available: https://arxiv.org/abs/2205.11096, 2022.

[54] T. N. Kipf, M. Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, 2017.

[55] Y. Ouali, C. Hudelot, M. Tami. An overview of deep semi-supervised learning, [Online], Available: https://arxiv.org/abs/2006.05278, 2020.

[56] Z. M. Zhang, Z. W. Yao, Y. Q. Yang, Y. J. Yan, J. E. Gonzalez, M. W. Mahoney. Benchmarking semi-supervised federated learning, [Online], Available: https://arxiv.org/abs/2008.11364v1, 2020.

[57] S. Itahara, T. Nishio, Y. Koda, M. Morikura, K. Yamamoto. Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-IID private data. *IEEE Transactions on Mobile Computing*, vol. 22, no. 1, pp. 191–205, 2023. DOI: 10.1109/TMC.2021.3070013.

[58] Y. Kang, Y. Liu, X. L. Liang. FedCVT: Semi-supervised vertical federated learning with cross-view training, [Online], Available: https://arxiv.org/abs/2008.10838, 2020.

[59] W. Jeong, J. Yoon, E. Yang, S. J. Hwang. Federated semi-supervised learning with inter-client consistency & disjoint learning. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[60] J. Xie, S. Y. Liu, J. X. Chen. A framework for distributed semi-supervised learning using single-layer feedforward networks. *Machine Intelligence Research*, vol. 19, no. 1, pp. 63–74, 2022. DOI: 10.1007/s11633-022-1315-6.

[61] K. Sohn, D. Berthelot, N. Carlini, Z. Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, C. L. Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, vol 33, pp. 596–608, 2020.

[62] Q. V. Le. Building high-level features using large scale unsupervised learning. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Vancouver, Canada, pp. 8595–8598, 2013. DOI: 10.1109/ICASSP.2013.6639343.

[63] F. D. Zhang, K. Kuang, Z. Y. You, T. Shen, J. Xiao, Y. Zhang, C. Wu, Y. T. Zhuang, X. L. Li. Federated unsupervised representation learning, [Online], Available: https://arxiv.org/abs/2010.08982, 2020.

[64] M. Servetnyk, C. C. Fung, Z. Han. Unsupervised federated learning for unbalanced data. In *Proceedings of IEEE Global Communications Conference*, IEEE, Taipei, China, 2020. DOI: 10.1109/GLOBECOM42002.2020.9348203.

[65] E. Tzinis, J. Casebeer, Z. P. Wang, P. Smaragdis. Separate but together: Unsupervised federated learning for speech enhancement from non-IID data. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, IEEE, New Paltz, USA, pp. 46–50, 2021. DOI: 10.1109/WASPAA52581.2021.9632783.

[66] W. Kim, A. Kanezaki, M. Tanaka. Unsupervised learning of image segmentation based on differentiable feature clustering. *IEEE Transactions on Image Processing* 29 (2020): 8055–8068. DOI: 10.1109/TIP.2020.3011269.

[67] Y. Q. Chen, X. Qin, J. D. Wang, C. H. Yu, W. Gao. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 83–93, 2020. DOI: 10.1109/MIS.2020.2988604.

[68] J. Y. Chen, A. D. Zhang. FedMSplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ACM, Washington DC, USA, pp. 87–96, 2022. DOI: 10.1145/3534678.3539384.

[69] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, T. Mikolov. DeViSE: A deep visual-semantic embedding model. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, ACM, Lake Tahoe, USA, pp. 2121–2129, 2013. DOI: 10.5555/2999792.2999849.

[70] J. E. Van Engelen, H. H. Hoos. A survey on semi-supervised learning. *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020. DOI: 10.1007/s10994-019-05855-6.

[71] Y. A. Chung, C. C. Wu, C. H. Shen, H. Y. Lee, L. S. Lee. Audio Word2Vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, Interspeech, San Francisco, USA, pp. 765–769, 2016.

[72] R. Zhang, P. Isola, A. A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 645–654, 2017. DOI: 10.1109/CVPR.2017.76.

[73] H. Choi, M. Kim, G. Lee, W. Kim. Unsupervised learning approach for network intrusion detection system using autoencoders. *The Journal of Supercomputing*, vol. 75, no. 9, pp. 5597–5621, 2019. DOI: 10.1007/s11227-019-02805-w.

[74] Y. C. Zhao, H. Y. Liu, H. L. Li, P. Barnaghi, H. Haddadi. Semi-supervised federated learning for activity recognition, [Online], Available: https://arxiv.org/abs/2011.00851, 2020.

[75] Y. C. Zhao, P. Barnaghi, H. Haddadi. Multimodal federated learning, [Online], Available: https://arxiv.org/abs/2109.04833v1, 2021.

[76] H. Z. Yu, Z. K. Chen, X. Zhang, X. Chen, F. Z. Zhuang, H. Xiong, X. Z. Cheng. FedHAR: Semi-supervised online learning for personalized federated human activity recognition. *IEEE Transactions on Mobile Computing*, published online. DOI: 10.1109/TMC.2021.3136853.

[77] A. Rahate, R. Walambe, S. Ramanna, K. Kotecha. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, vol. 81, pp. 203–239, 2022. DOI: 10.1016/j.inffus.2021.12.003.

[78] J. B. Wang, G. Y. Xie, Y. W. Huang, Y. F. Zheng, Y. C. Jin, F. Zheng. FedMed-ATL: Misaligned unpaired brain image synthesis via affine transform loss, [Online], Available: https://arxiv.org/abs/2201.12589, 2022.

[79] A. Saeed, F. D. Salim, T. Ozcelebi, J. Lukkien. Federated self-supervised learning of multisensor representations for embedded intelligence. *IEEE Internet of Things Journal*, vol. 8, no. 2, pp. 1030–1040, 2021. DOI: 10.1109/JIOT.2020.3009358.

[80] K. S. Arikumar, S. B. Prathiba, M. Alazab, T. R. Gadekallu, S. Pandya, J. M. Khan, R. S. Moorthy. FL-PMI: Federated learning-based person movement identification through wearable devices in smart healthcare systems. *Sensors*, vol. 22, no. 4, Article number 1377, 2022. DOI: 10.3390/s22041377.

[81] Y. L. Sun. Federated Transfer Learning with Multimodal Data, [Online], Available: https://arxiv.org/abs/2209.03137, 2022.

[82] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman. The kinetics human action video dataset, [Online], Available: https://arxiv.org/abs/1705.06950, 2017.

[83] T. Sztyler, H. Stuckenschmidt. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *Proceedings of IEEE International Conference on Pervasive Computing and Communications*, IEEE, Sydney, Australia, pp. 1–9, 2016. DOI: 10.1109/PERCOM.2016.7456521.

[84] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008. DOI: 10.1007/s10579-008-9076-6.

[85] Z. R. Wu, S. R. Song, A. Khosla, F. Yu, L. G. Zhang, X. O. Tang, J. X. Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 1912–1920, 2015. DOI: 10.1109/CVPR.2015.7298801.

[86] M. F. Duarte, Y. H. Hu. Vehicle classification in distributed sensor networks. *Journal of Parallel and Distributed Computing*, vol. 64, no. 7, pp. 826–838, 2004. DOI: 10.1016/j.jpdc.2004.03.020.

[87] A. Banos, R. Garcia, J. A. Holgado-Terriza, M. Damas, H. Pomares, I. Rojas, A. Saez, C. Villalonga. mHealthDroid: A novel framework for agile development of mobile health applications. In *Proceedings of the 6th International Workshop on Ambient Assisted Living and Daily Activities*, Springer, Belfast, UK, pp. 91–98, 2014. DOI: 10.1007/978-3-319-13105-4_14.

[88] B. Kwolek, M. Kepski. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer Methods and Programs in Biomedicine*, vol. 117, no. 3, pp. 489–501, 2014. DOI: 10.1016/j.cmpb.2014.09.005.

[89] J. Cañedo, A. Skjellum. Using machine learning to secure IoT systems. In *Proceedings of the 14th Annual Conference on Privacy, Security and Trust*, IEEE, Auckland, New Zealand, pp. 219–222, 2016. DOI: 10.1109/PST.2016.7906930.

[90] L. Xiao, X. Y. Wan, X. Z. Lu, Y. Y. Zhang, D. Wu. IoT security techniques based on machine learning: How do iot devices use ai to enhance security? *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 41–49, 2018. DOI: 10.1109/MSP.2018.2825478.

[91] F. Zantalis, G. Koulouras, S. Karabetsos, D. Kandris. A review of machine learning and iot in smart transportation. *Future Internet*, vol. 11, no. 4, Article number 94, 2019. DOI: 10.3390/fi11040094.

[92] A. Franco, A. Magnani, D. Maio. A multimodal approach for human activity recognition based on skeleton and RGB data. *Pattern Recognition Letters*, vol, 131, pp. 293–299, 2020. DOI: 10.1016/j.patrec.2020.01.010.

[93] Z. Q. Zhu, S. Wan, P. Y. Fan, K. B. Letaief. Federated multiagent actor–critic learning for age sensitive mobile-edge computing. *IEEE Internet of Things Journal*, vol. 9, no. 2, pp. 1053–1067, 2022. DOI: 10.1109/JIOT.2021.3078514.

[94] S. Wang, C. Y. Li, D. W. K. Ng, Y. C. Eldar, H. V. Poor, Q. Hao, C. Z. Xu. Federated deep learning meets autonomous vehicle perception: Design and verification. IEEE Network, 2022. (Online first). DOI: 10.1109/MNET.104.2100403.

[95] S. J. Chen, B. C. Li. Towards optimal multi-modal federated learning on non-IID data with hierarchical gradient blending. In *Proceedings of IEEE Conference on Computer Communications*, IEEE, London, UK, pp. 1469–1478, 2022. DOI: 10.1109/INFOCOM48880.2022.9796724.

[96] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, F. Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, vol. 5, no. 1, pp. 1–19, 2021. DOI: 10.1007/s41666-020-00082-4.

[97] X. H. He, X. Y. Yang, S. H. Zhang, J. Y. Zhao, Y. C. Zhang, E. Xing, P. T. Xie. Sample-efficient deep learning for COVID-19 diagnosis based on CT scans. *medRxiv*, published online. DOI: 10.1101/2020.04.13.20063941.

[98] A. M. Ismael, A. Şengür. Deep learning approaches for COVID-19 detection based on chest X-ray images. *Expert Systems with Applications*, vol. 164, Article number 114054, 2021. DOI: 10.1016/j.eswa.2020.114054.

[99] K. Gong, D. F. Wu, C. D. Arru, F. Homayounieh, N. Neumark, J. H. Guan, V. Buch, K. Kim, B. C. Bizzo, H. Ren, W. Y. Tak, S. Y. Park, Y. R. Lee, M. K. Kang, J. G. Park, A. Carriero, L. Saba, M. Masjedi, H. Talari, R. Babaei, H. K. Mobin, S. Ebrahimian, N. Guo, S. R. Digumarthy, I. Dayan, M. K. Kalra, Q. Z. Li. A multicenter study of COVID-19 patient prognosis using deep learning-based CT image analysis and electronic health records. *European Journal of Radiology*, vol. 139, Article number 109583, 2021. DOI: 10.1016/j.ejrad.2021.109583.

[100] V. S. Parekh, S. H. Lai, V. Braverman, J. Leal, S. Rowe, J. J. Pillai, M. A. Jacobs. Cross-domain federated learning in medical imaging, [Online], Available: https://arxiv.org/abs/2112.10001, 2021.

[101] X. M. Chen, Y. X. Shao, Z. Xue, Z. Q. Yu. Multi-modal COVID-19 discovery with collaborative federated learning. In *Proceedings of the 7th International Conference on Cloud Computing and Intelligent Systems*, IEEE, Xi′an, China, pp. 52–56, 2021. DOI: 10.1109/CCIS53392.2021.9754623.

[102] J. Ji, D. F. Yan, Z. Y. Mu. Personnel status detection model suitable for vertical federated learning structure. In *Proceedings of the 6th International Conference on Machine Learning and Soft Computing*, ACM, Haikou, China, pp. 98–104, 2022. DOI: 10.1145/3523150.3523166.

[103] T. Baltrušaitis, C. Ahuja, L. P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019. DOI: 10.1109/TPAMI.2018.2798607.

[104] T. Baltrušaitis, C. Ahuja, L. P. Morency. Challenges and applications in multimodal machine learning. *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition*, S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, A. Krüger, Eds., ACM, pp. 17–48, 2018. DOI: 10.1145/3107990.3107993.

[105] Q. Chang, H. Qu, Z. N. Yan, Y. H. Gao, L. Baskaran, D. Metaxas. Modality bank: Learn multi-modality images across data centers without sharing medical data. In *Proceedings of the 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, IEEE, Glasgow, UK, pp. 4758–4763, 2022. DOI: 10.1109/EMBC48229.2022.9871529.

[106] C. Wang, M. Niepert, H. Li. LRMM: Learning to recommend with missing modalities. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Brussels, Belgium, pp. 3360–3370, 2018. DOI: 10.18653/v1/D18-1373.

[107] Y. Shen, M. C. Gao. Brain tumor segmentation on mri with missing modalities. In *Proceedings of the 26th International Conference on Information Processing in Medical Imaging*, Springer, Hong Kong, China, pp. 417–428, 2019. DOI: 10.1007/978-3-030-20351-1_32.

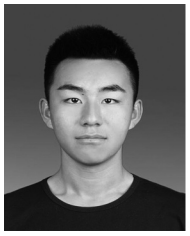
[108] F. Ma, S. L. Huang, L. Zhang. An efficient approach for audio-visual emotion recognition with missing labels and missing modalities. In *Proceedings of IEEE International Conference on Multimedia and Expo*, IEEE, Shenzhen, China, pp. 1–6, 2021. DOI: 10.1109/ICME51207.2021.9428219.

[109] A. Sadilek, L. Y. Liu, D. Nguyen, M. Kamruzzaman, S. Serghiou, B. Rader, A. Ingerman, S. Mellem, P. Kairouz, E. O. Nsoesie, J. Macfarlane, A. Vullikanti, M. Marathe, P. Eastham, J. S. Brownstein, B. A. Y. Arcas, M. D. Howell, J. Hernandez. Privacy-first health research with federated learning. *NPJ Digital Medicine*, vol. 4, no. 1,

Article number 132, 2021. DOI: 10.1038/s41746-021-00489-2.

[110] I. Balelli, S. Silva, M. Lorenzi. A differentially private probabilistic framework for modeling the variability across federated datasets of heterogeneous multi-view observations, [Online], Available: https://arxiv.org/abs/2204.07352, 2022.

[111] L. Zhang, W. Cui, B. Li, Z. H. Chen, M. Wu, T. S. Gee. Privacy-preserving cross-environment human activity recognition. *IEEE Transactions on Cybernetics*, vol. 53, no. 3, pp. 1765–1775, 2023. DOI: 10.1109/TCYB.2021.3126831.



**Yi-Ming Lin** is currently a master student in electronic science and technology from Key Laboratory of Collaborative Intelligence Systems, Ministry of Education, Xidian University, China.

His research interests include federated learning and multimodal learning.

E-mail: 1169086366@qq.com

ORCID iD: 0000-0002-1535-427X

**Yuan Gao** received the B. Eng. degree in intelligent science and technology from Xidian University, China in 2018. He received the Ph. D. degree in pattern recognition and intelligent systems from Xidian University, China in 2022. He is currently a lecturer with Key Laboratory of Collaborative Intelligence Systems, Ministry of Education, Xidian University, China.

His research interests include edge computing, secure artificial intelligence and collaborative multiparty learning.

E-mail: cn_gaoyuan@foxmail.com

ORCID iD: 0000-0002-2990-9205

**Mao-Guo Gong** received the B. Sc. degree in electronic engineering (first class honors) and the Ph. D. degree in electronic science and technology from Xidian University, China in 2003 and 2009, respectively. Since 2006, he has been a teacher with Xidian University. In 2008 and 2010, he was promoted as an associate professor and as a full professor, respectively, both with exceptional admission. He received the prestigious National Program for the support of Top-Notch Young Professionals from the Central Organization Department of China, the Excellent Young Scientist Foundation from the National Natural Science Foundation of China, and the New Century Excellent Talent in University from the Ministry of Education of China. He is also an Associate Editor of *IEEE Transactions on Evolutionary Computation* and *IEEE Transactions on Neural Networks and Learning Systems*.

His research interests include computational intelligence with applications to optimization, learning, data mining and image understanding.

E-mail: gong@ieee.org (Corresponding author)

ORCID iD: 0000-0002-0415-8556

**Si-Jia Zhang** is currently a Ph. D. degree candidate in pattern recognition and intelligent systems from Key Laboratory of Collaborative Intelligence Systems, Ministry of Education, Xidian University, China.

Her research interests include deep learning and image processing.

E-mail: zsj_stella@foxmail.com

**Yuan-Qiao Zhang** received the B. Eng. degree in intelligent science and technology from School of Electronic Engineering, Xidian University, China in 2016. He is currently a Ph. D. degree candidate in electronic science and technology from Key Laboratory of Collaborative Intelligence Systems, Ministry of Education, Xidian University, China.

His research interests include multi-objective evolutionary optimization and privacy protection.

E-mail: zhangyq@stu.xidian.edu.cn

**Zhi-Yuan Li** received a B. Eng. degree in intelligent science and technology from School of Electronic and Information Engineering, Xi′an Jiaotong University, China, and the M. Eng. degree in computer science and technology from Tandon School of Engineering, New York University. Currently, he is a Ph. D. degree candidate in electronic science and technology from School of Electronic Engineering, Xidian University, China.

His research interests include intelligent computing and secure AI.

E-mail: zhiyuanli2022@gmail.com